Pharmacoecon Ger Res Artic (2013) 11:101–117
DOI 10.1007/s40275-014-0012-7

ÜBERSICHTSARTIKEL

Experimentelle Präferenzmessung im Gesundheitswesen mit Hilfe von Best-Worst Scaling (BWS)

Axel C. Mühlbacher · Anika Kaczynski · Peter Zweifel

Online publiziert: 2. August 2014

Abstract Best-Worst Scaling (BWS) is a method of multi-attribute preference measurement. Its objective is to determine the preferences with respect to certain properties and characteristics. It is a stated preference method and based on the assumption that people are able to select the best and worst or subjectively most and least important from a set of three or more elements. BWS avoids, like all discrete choice experiments, the known weaknesses of rating and ranking scales. However, BWS promises to generate additional information because participants choose two times, namely the best as well as the worst alternative. This paper describes the potentials of application, the underlying theoretical concepts, and the implementation of the three variants of BWS experiments. It also shows that at least the second variant of BWS (the so-called profile case) is subject to restrictive and unrealistic assumptions.

1 Präferenzen im Gesundheitswesen

Ein zentrales Problem der Entscheidungsträger im Gesundheitswesen ist die optimale Allokation der vorhandenen Ressourcen. Unabhängig davon, ob die Entscheidungen auf regulatorischer oder auf der medizinischen Ebene den Ressourceneinsatz beeinflussen, sind die Bürgerinnen und Bürger betroffen, welche ihrerseits bei ihren Entscheidungen optimale (d. h. für sie bestmögliche) Lösungen suchen. Das Wort „bestmöglich“ deutet bereits auf die zentrale Rolle individueller Bewertungen und damit individueller, subjektiver Präferenzen hin. Wenn demnach die Entscheidungsträger im Gesundheitswesen die Ressourcen in einer Art und Weise einsetzen, die nicht mit den Präferenzen der Betroffenen übereinstimmt, kann von einer ineffizienten Nutzung der vorhandenen knappen Ressourcen ausgegangen werden [1, 2]. Gesundheitspolitische Entscheidungsträger verfehlen ihr Ziel, denn die Interventionen maximieren nicht den Nutzen der Versicherten und Patienten, sind also nicht optimal aus der Perspektive der Versicherten oder Bürger und leiden unter mangelnder Akzeptanz der Betroffenen. Gemäss der sog. Public Choice Theorie [2], die davon ausgeht, dass Politiker und Beamte ihre eigenen Ziele verfolgen, ist dies ein realistisches Szenario.

Ein Beispiel ist die Versorgung von Diabetespatienten. Viele Diabetiker sträuben sich anfangs gegen eine Insulintherapie, nicht zuletzt aus Angst vor einer möglichen Gewichtszunahme (mehrheitlich wird eine Gewichtsabnahme favorisiert). Die Zunahme des Gewichts ist eine Folge des verbesserten Stoffwechsels, denn je niedriger der Langzeit-Blutzuckerspiegel desto mehr steigt das Gewicht [3]. Für die Patienten stellt sich die Frage, ob die Verbesserung des Stoffwechsels das Risiko der Gewichtszunahme wert ist. Besonders wenn sie keine Erfahrungen mit Hypoglykämien (Unterzuckerungen) haben, könnten oral therapierte Patienten (OAD) im Unterschied zu insulinerfahrenen Patienten dazu nicht bereit sein [4, 6].

Dieses Beispiel illustriert die Bedeutung, welche der Erfassung individueller Präferenzen gerade im Gesundheitswesen zukommt. Denn im Gegensatz zu anderen Sektoren der Wirtschaft, wo die Konsumenten ihren Präferenzen unmittelbar durch die individuelle Wahlleistung (den Kauf oder eben Nichtkauf von Gütern und Leistungen) Ausdruck

A.C. Mühlbacher (✉) · A. Kaczynski
IGM Institut Gesundheitsökonomie und Medizinmanagement,
Hochschule Neubrandenburg, Brodaer Straße 2,
17033 Neubrandenburg, Deutschland
e-mail: muehlbacher@hs-nb.de

P. Zweifel
Universität Zürich, Kreuth 371, 9531 Bad Bleiberg, Österreich
e-mail: peter.zweifel@econ.uzh.ch

geben können, ist im Gesundheitswesen die Konsumentensouveränität im Wesentlichen auf die Wahl der Krankenversicherung beschränkt [5]. Im Krankheitsfall entscheiden nach wie vor die Ärzte und andere Vertreter der Gesundheitsberufe über die Bereitstellung und Inanspruchnahme von Gesundheitsleistungen. Andererseits soll die Partizipation der Patienten bei medizinischen Entscheidungen gefördert werden, um die zur Verfügung stehenden Ressourcen gezielter einzusetzen und die Versorgungsleistungen zu optimieren. Doch nach wie vor müssen Entscheidungen hinsichtlich zukünftiger Versorgungsstrategien die Präferenzen aller Stakeholder im Gesundheitswesen berücksichtigen.

Aufgrund der enormen Bedeutung der Präferenzen werden in der Versorgungsforschung zunehmend experimentelle Methoden zu ihrer Messung eingesetzt. Allerdings erlauben nicht alle Methoden, die Präferenzen einer bestimmten Gruppe richtig zu erfassen und sie von denjenigen einer anderen Gruppe klar zu unterscheiden. Dieser Beitrag verfolgt deshalb zwei Ziele. Zum einen soll er einen Überblick über die verbreiteten Methoden (Rating Scale, Ranking, Best-Worst Scaling (BWS)) verschaffen. Dazu gehören auch das experimentelle Design sowie die statistischen Inferenzverfahren. Zum anderen geht es darum, die Schwächen der verbreiteten Methoden im Lichte einer nutzentheoretischen Fundierung aufzuzeigen. In diesem Lichte erscheinen sog. Discrete-Choice Experimente mit festem Status quo, gegebenenfalls in Kombination mit BWS, als die überzeugendste Alternative [7]. Sie ermöglichen, die relative Wichtigkeit der Attribute auch gruppenspezifisch zu messen und sie zu einem Gesamtnutzenwert zu aggregieren, der auch als monetärer Wert ausgedrückt und damit den Kosten gegenübergestellt werden kann.

2 Übersicht über die Methode

2.1 Mikroökonomische Grundlagen

Präferenzen und Indifferenzkurve Die mikroökonomische Erklärung von Präferenzen basiert auf der Annahme, dass rationale Individuen (allgemeiner: Entscheidungsträger) stets diejenige Menge von Attributen mit ihren jeweiligen Ausprägungen bevorzugen, die ihren individuellen Nutzen maximiert [8]. Die Nutzenmaximierung bildet die Basis der Präferenz- und Nutzenmessung auch im Gesundheitswesen, auch wenn gerade hier gewisse Annahmen über die Präferenzordnung (Vollständigkeit, Reflexivität und Transitivität [9]) kritisch beurteilt werden. Stehen mehrere Alternativen zur Auswahl, entscheiden sich die Individuen annahmegemäß für jene Alternative, welche aufgrund ihrer Eigenschaften (Attribute) den höchsten Grad an Bedürfnisbefriedigung erwarten lässt. Die Präferenzen werden durch sog. Indifferenzkurven abgebildet (vgl. Abb. 1). Dabei soll das eingangs zitierte Beispiel des Abwägens zwischen der Wirkung (Senkung des Langzeit-Zuckerwertes) und der Nebenwirkung (Gewichtszunahme respektive Gewichtsreduktion) durch Insulinpatienten und OAD-Patienten aufgegriffen werden.

Der Status quo (Punkt S) zeigt die konventionelle Versorgung; er sei durch zwei Attribute gekennzeichnet, nämlich ein Niveau des Langzeit-Zuckerwertes (a_1) gepaart mit einem bestimmten Grad der Gewichtsabnahme (a_2). Zwei sog. Indifferenzkurven verlaufen durch S. Sie zeigen jene Kombinationen von Mengen (oder Ausprägungen) der beiden Attribute an, die vom OAD-Patienten (bzw. Insulinnaiven Diabetiker) als gleichwertig eingestuft werden, d. h. den gleichen Nutzen haben. Die Indifferenzkurve des Insulinpatienten verläuft steil. Um das genannte Beispiel zum Ausdruck zu bringen, würde der Insulinpatient für eine Senkung des Langzeit-Zuckerwertes (Δa_1) eine deutlich höhere Gewichtsreduktion Δa_2 in Kauf nehmen, ohne einen Nutzenverlust zu erleiden. Allgemein zeigt demnach die Steigung der Indifferenzkurve die relative Wichtigkeit eines Attributs und damit die Präferenzstruktur aus der Sicht der betroffenen Person an.

Die Bestimmung der Indifferenzkurve Die experimentelle Präferenzermittlung lässt sich für den OAD-Patienten wie folgt zeigen. Ein Proband wird gebeten, die Kombination der Eigenschaften X (mehr von Attribut a_2, dafür deutlich weniger von Attribut a_1) mit dem Status quo S zu vergleichen. Wählt er oder sie X, ist X besser als S, d. h. die (unbekannte) Indifferenzkurve muss unterhalb von X verlaufen. Jetzt werden die Attribute durch den Experimentator neu gemischt, und der Proband wird gebeten, die Kombination Y gegen den Status quo S abzuwägen. Gibt er oder sie jetzt S den Vorzug, so verläuft die Indifferenzkurve oberhalb von Y. Durch Wiederholung dieses Vorgangs (z. B. mit den Punkten B[1] und W[2], die weiter unten von Bedeutung sein werden) wird es möglich, die Indifferenzkurve zu schätzen. Deren Steigung $\Delta a_2^{\mathrm{OAD}}/\Delta a_1$ zeigt an, dass für den OAD-Patienten eine Senkung des Langzeit-Zuckerwertes durch eine nur relativ geringe Gewichtsreduktion aufgewogen werden könnte. Ganz anders beim Insulinpatienten: Dort ist das Verhältnis $\Delta a_2^I/\Delta a_1$ sehr viel höher und besagt, dass Insulinpatienten für eine Senkung des Langzeit-Zuckerwertes deutlich höher kompensiert werden müssten.

Die Bestimmung monetärer Äquivalente Es ist sogar möglich, mit einem DCE den (subjektiven) Wert eines Attributs in Geld auszudrücken. Dafür genügt es, in der Abb. 1 den Langzeit-Zuckerwert durch das Einkommen zu ersetzen, das mit einem Abweichen vom Status quo verbunden ist. Dann misst die Steigung der Indifferenzkurve $\Delta a_2^{\mathrm{OAD}}/\Delta a_1$ die (marginale) Zahlungsbereitschaft (MZB, engl. willingness to pay) des betrachteten Probanden für eine kleine Reduktion

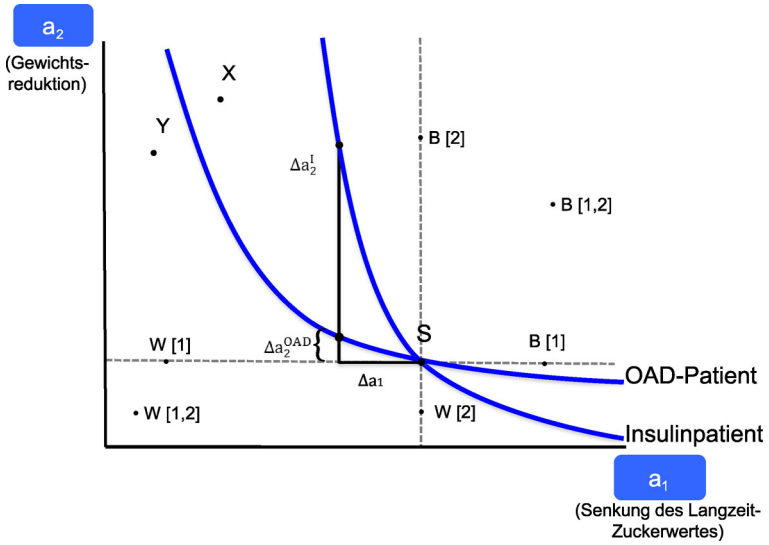

Abb. 1 Präferenzermittlung mit DCE

on des Gewichtes. Schließlich lässt sich der in Geld ausgedrückte Gesamtnutzen eines Eigenschaftenbündels aus den MZB-Werten der einzelnen Attribute bestimmen, wobei allerdings nur dann ein globaler Wert unmittelbar bestimmt werden kann, wenn die postulierte Nutzenfunktion linear ist und keine Interaktionsterme aufweist [10, 11]. Ein DCE erlaubt demnach die Bestimmung sowohl von Teilnutzen- wie auch Gesamtnutzenwerten [12].

Interpretation der Ergebnisse unabhängig von der Form der Nutzenfunktion Ein erster Vorteil des Arbeitens mit der Indifferenzkurve besteht darin, dass ihre Steigung unabhängig von der Form der Nutzenfunktion ist, d. h. die MZB hängt nicht von der Wahl der Funktion ab, mit der die Attribute bewertet werden. Dies lässt sich wie folgt zeigen: Es sei $U(a_1, a_2) = u$ die Nutzenfunktion, welche die subjektive Bewertung einer Eigenschaftskombination angibt. Entlang einer Indifferenzkurve ist der Nutzen definitionsgemäß konstant; anders ausgedrückt: die Veränderung des Nutzens ΔU muss Null sein. Eine solche Veränderung könnte an sich auch auf eine Veränderung der Funktion $f(\cdot)$ zurückgehen, doch dies würde die Wahlhandlungen der Probanden jeder Konsistenz berauben. Ein ΔU kann dann aber nur durch Veränderungen in den Ausprägungen der Eigenschaften Δa_1 und Δa_2 zustande kommen. Es gilt also

$$\Delta U = \frac{\partial f}{\partial a_1} \cdot \Delta a_1 + \frac{\partial f}{\partial a_2} \cdot \Delta a_2, \tag{1}$$

wobei $\partial f/\partial a_1$ und $\partial f/\partial a_2$ die Bedeutung des betreffenden Attributs für den Nutzen der betrachteten Person (Grenznutzen) anzeigt. Nullsetzen von ΔU und Auflösen nach $\Delta a_2/\Delta a_1$ ergibt die in Abb. 1 eingetragene Steigung der Indifferenzkurve:

$$\frac{\Delta a_2}{\Delta a_1} = \frac{\partial f/\partial a_2}{\partial f/\partial a_1} \tag{2}$$

Jetzt soll die Nutzenfunktion einer beliebigen Transformation $\varphi(\cdot)$ unterworfen werden. Mit $\tilde{U} = \varphi(U)$ lautet dann die Gleichung für die neue Indifferenzkurve

$$\Delta\tilde{U} = \frac{\partial \varphi}{\partial f} \cdot \frac{\partial f}{\partial a_1} \cdot \Delta a_1 + \frac{\partial \varphi}{\partial f} \cdot \frac{\partial f}{\partial a_2} \cdot \Delta a_2 = 0. \tag{3}$$

Wenn man diese Gleichung wiederum für die Steigung $\Delta a_2/\Delta a_1$ auflöst, erhält man

$$\frac{\Delta a_2}{\Delta a_1} = -\frac{\partial \varphi/\partial f \cdot \partial f/\partial a_2}{\partial \varphi/\partial f \cdot \partial f/\partial a_1} = -\frac{\partial f/\partial a_2}{\partial f/\partial a_1}. \tag{4}$$

Mit anderen Worten, die Steigung der Indifferenzkurve wird von der Wahl der Nutzenfunktion nicht berührt.

Homothetik für Skaleninvarianz Um die Skaleninvarianz bezüglich der Argumente a_1 und a_2 zu gewährleisten, braucht es die Annahme der sog. Homothetik, d. h. die Funktion $\tilde{U} = \varphi(U)$ muss die Form $\tilde{U} = g(\lambda) \cdot U$ haben, wobei $g(\lambda)$ eine monoton ansteigende Funktion eines gemeinsamen Skalierungsfaktors $\lambda > 0$ angewendet auf die beiden Attribute darstellt. Somit gilt bei Homothetik

$$\tilde{U} = g(\lambda) \cdot U = f(\lambda a_1, \lambda a_2). \tag{5}$$

Die Gleichung für die Indifferenzkurve auf dem Niveau λ lautet dann

$$\Delta\{g(\lambda) \cdot U\} = \frac{\partial f}{\partial(\lambda a_1)} \cdot \Delta(\lambda a_1) + \frac{\partial f}{\partial(\lambda a_2)} \cdot \Delta(\lambda a_2) = 0 \tag{6}$$

mit Steigung

$$\frac{\Delta(\lambda a_2)}{\Delta(\lambda a_1)} = -\frac{\partial f/\partial(\lambda a_2)}{\partial f/\partial(\lambda a_1)}. \tag{7}$$

Da λ eine feste Größe ist, kann es auf der linken Seite der (7) ausgeklammert werden und hebt sich weg. Was die rechte

Seite betrifft, so ergibt die partielle Ableitung

$$\partial f/\partial(\lambda a_1) = \partial f/\partial a_1 \cdot \left[\partial a_1/\partial(\lambda a_1)\right] = (1/\lambda) \cdot \partial f/\partial a_1$$

und analog

$$\partial f/\partial(\lambda a_2) = (1/\lambda) \cdot \partial f/\partial a_2. \tag{8}$$

Der Faktor $(1/\lambda)$ hebt sich auf der rechten Seite weg, so dass man für die Steigung der Indifferenzkurve auf dem Niveau λ nach der Skalierung erhält

$$\frac{\Delta(a_2)}{\Delta(a_1)} = -\frac{\partial(f)/\partial(a_2)}{\partial(f)/\partial(a_1)}, \tag{9}$$

was mit der (4) identisch ist. Da durch die Skalierung das Verhältnis a_2/a_1 nicht verändert wird, bleibt entlang eines Fahrstrahls durch den Ursprung der Abb. 1 die Steigung der Indifferenzkurve bei Homothetik konstant; jede Indifferenzkurve ist die Kopie ihrer Nachbarn. Damit wird die relative Wertung sowie die MZB der Attribute Skalen-invariant.

Erhebung individueller Präferenzen Die Erhebung der Präferenzen über die Wahlhandlungen steht im Wettbewerb mit zwei anderen Verfahren, nämlich dem Rating (bei dem die Befragten die Alternativen mit Zahlenwerten versehen) und dem Ranking (bei dem sie eine Reihung der Alternativen vornehmen). Es existieren demnach drei experimentelle Erhebungsmethoden, mit deren Hilfe die Präferenzen gemessen werden können: Rating (Bewertung), Ranking (Reihung) und Choice (Wahlentscheidung) [13, 14]. Aus nutzentheoretischer Sicht kommt die Wahl (Choice) einer tatsächlichen Entscheidung am nächsten (entweder kauft man ein Produkt oder lässt es sein), obschon in einem Experiment die teilnehmenden Probanden – im Gegensatz zur Realität – die Konsequenzen ihrer Wahlhandlungen nicht zu spüren bekommen (z. B. da sie einen Kaufpreis nicht entrichten müssen) [14].

Präferenzdaten auf Basis von Bewertungen Beim Rating-Verfahren erfolgt die Einstufung des Gesamtnutzens auf einer metrischen Skala [15, 16]. Die sog. Rating Scale besteht zumeist aus einer Geraden mit eindeutig definierten Endpunkten, wobei die Endpunkte die höchsten und niedrigsten Bewertungen darstellen [17]. Weitere Formen von Ratingskalen sind z. B. die Likert-Skalen oder die kontinuierlichen Analogskalen [18]. Dieses Präferenzmaß ist relativ einfach zu erheben, weist aber methodische Schwächen auf. Erstens neigen die Befragten beim Rating dazu, wenig diskriminierende Bewertungen abzugeben, was zu einer mangelnden Differenzierung der Antworten führt. Sie können auch viele oder gar alle Alternativen gleich bewerten (sog. Deckeneffekte); mit Blick auf die Abb. 1 liegen die Alternativen auf derselben Indifferenzkurve. Zudem werden die Antworten häufig von sozialen Erwartungen beeinflusst (Ja-Sage- bzw. Nein-Sage-Tendenz). Im einführenden Beispiel könnten Insulinpatienten unter dem Eindruck stehen, eine Gewichtsreduktion werde allgemein begrüßt. Beim Rating ordnen sie den Alternativen zu hohe Nutzenwerte zu [19]. Rating-Skalen stellen hohe Anforderungen an die Befragten, da sie ihre Präferenzen auf die vorgegebene Skala übertragen müssen. Zudem sind sie äußerst anfällig für inkonsistente und zufällige Antworten. Unklar ist hierbei auch, ob und inwieweit die Probanden die verschiedenen Attribute tatsächlich in Relation zueinander bewerten, wie dies aus der Steigung der Indifferenzkurve hervorgeht [16]. Somit ist die Konsistenz der Antworten bei diesem Frageformat nicht immer zufriedenstellend [20].

Aus der Sicht der (mikroökonomischen) Nutzentheorie besteht der entscheidende Nachteil jedoch darin, dass das Rating-Verfahren nicht Skalen-invariant ist. Ein Unterschied z. B. zwischen 50 und 60 Punkten auf der Skala kann je nach Proband eine ganz unterschiedliche Nutzendifferenz abbilden. In der Abb. 1 sollen diese zehn Punkte dem Abstand zwischen den Punkten X und Y entsprechen. Für den OAD-Patienten mag diese Differenz einen großen Nutzenunterschied bedeuten. Für den Insulinpatienten kann diese Differenz dagegen einen geringen Unterschied bedeuten, denn beide Punkte liegen weit weg von seiner (unbekannten) Indifferenzkurve durch den Status quo S. Die Abhängigkeit von der Skalierung führt auch dazu, dass sich die relative Bedeutung der Eigenschaften (und damit die Präferenzstruktur) nicht ermitteln lässt. Wenn Attribut a_1 z. B. mit Nutzenwerten 10, 20, 30, a_2 dagegen mit 20, 40, 60 abgebildet wird, erscheint Attribut a_2 doppelt so wichtig wie a_1. Man hätte dem Attribut a_2 aber ebenso gut die gleichen Werte 10, 20, 30 wie a_1 zuordnen können, mit der Folge, dass es als gleich wichtig erscheinen würde. Unabhängig von der Skalierung würde der Experimentator jedoch feststellen, dass die Probanden stets der dritten Alternative den Vorzug geben; er kann demnach nicht von den beobachtbaren Entscheidungen auf die richtige Skala schließen. Die mit Hilfe der Rating-Skalen ermittelten Präferenzen werden deshalb in der Regel verfälscht dargestellt und erlauben keine Rückschlüsse auf Präferenzunterschiede zwischen Individuen.

Präferenzdaten auf Basis von Reihungen Die Ranking-Methode verwendet ein nicht-metrisches Präferenzmaß, indem die Probanden gebeten werden, die aufgeführten Alternativen in eine Rangordnung zu bringen. Eine methodische Stärke dieses Verfahrens ist die höhere Validität und Reliabilität der Urteile im Vergleich zum Rating-Verfahren, denn die Probanden müssen hier ihre Präferenzen nicht in einer Weise zum Ausdruck bringen, dass der Unterschied zwischen 80 und 50 Punkten dem Dreifachen des Unterschieds zwischen 60 und 50 Punkten entspricht. Die Angabe, dass die Alternative ‚80' besser ist als die Alternative ‚60' und

diese wieder besser als ‚50', genügt [16, 21]. Immerhin werden sie angehalten, eine vollständige Präferenzordnung zu bilden, was ein Abwägen der verschiedenen Attribute gegeneinander verlangt [22].

Das Ranking-Verfahren steht im Einklang mit der Nutzentheorie, hat aber auch seine Schwächen. So bedingt es einen viel höheren umfragetechnischen Aufwand als das DCE, dem auf Seiten der Probanden eine potentielle kognitive Überforderung und Überlastung entspricht [16]. Um diese zu verhindern, muss die Zahl der zu rangierenden Alternativen klein gehalten werden, was auch eine geringe Zahl von Attributen verlangt (sonst müssen den Probanden sehr viel Alternativen vorgelegt werden, damit ein Attribut mindestens einmal in einer Kombination vorkommt). Zudem werden eindeutige Rangordnungen verlangt, sonst können die zugrundeliegenden Präferenzen nicht adäquat abgebildet werden [20]. Die ordinale Skalierung der gewonnen Daten erschwert schließlich die Vergleichbarkeit und weiterführende Analysen [23].

Präferenzdaten auf Basis von Wahlhandlungen Die Erhebung von Wahlentscheidungen (Choice) steht der Nutzentheorie am nächsten, da sie keine weitergehenden als die oben beschriebenen Annahmen verlangt (vollständige, reflexive und transitive Präferenzordnung [9]). In Abb. 1 wird das sog. „Single Multiple Choice Model" (SMC), das dem klassischen Discrete-Choice Experiment gleicht, dargestellt, bei dem die Befragten aus den vorgelegten Alternativen nur jeweils eine auswählen. Es lassen sich so keine Informationen über die Beziehungen zwischen den nicht gewählten Alternativen gewinnen. Beim „Modified Multiple Choice"-Model (MMC) dagegen haben die Befragten die Möglichkeit, die zwei (oder auch drei und mehr) für sie wichtigsten (d. h. mit dem größten Grenznutzen) Attribute auszuwählen, was zusätzliche Information über ihre Präferenzen generiert. Während dieser Ansatz eine realitätsnahe Methode mit geringen kognitiven Hürden darstellt, erlaubt er es nicht, die Nutzenunterschiede sowohl der nicht gewählten wie auch der gewählten Alternativen zu bestimmen. Dies würde die Aussage „der Abstand zwischen Rang 1 und 3 ist doppelt so groß wie derjenige zwischen Rang 2 und 3" bedingen [20]. Als eine Sonderform der „Choice" kann „Paired Comparison" gelten. Bei „Paired Comparison" handelt es sich um eine Kombination aus Choice-Experiment (Wahl der am meisten präferierten Alternative) und Rating (Bewertung der Stärke der Präferenz). So erfolgt bei der „Choice" ausschließlich die Wahl der besten Alternativen, während bei „Paired Comparison" zusätzlich noch eine Abstufung in der Bewertung erfolgt [14]. Da es sich beim BWS um eine Form des Discrete-Choice Experimentes handelt, basieren die hier ermittelten Präferenzdaten ebenfalls auf Wahlentscheidungen („Choice").

2.2 Best-Worst Scaling

Verschiedene Autoren wiesen schon früh darauf hin, dass die oben dargestellten Ansätze zur Präferenzmessung teils hohe Anforderungen an die Urteilsfähigkeit der Probanden stellen, teils messtheoretische Probleme aufwerfen [24–26]. Als Antwort darauf wurde Ende der 1980er Jahre mit dem Best-Worst Scaling (BWS) die „Choice"-Methode weiterentwickelt. Es handelt sich dabei um eine Erweiterung der paarweisen Vergleiche, welche je nach Anwendung die Problemfelder der bislang verwendeten Methoden umgeht. Es werden drei Varianten (Cases) von BWS unterschieden [27]. Sie haben alle gemeinsam, dass die Befragten nicht nur die bessere Alternative wählen, sondern jeweils die „beste" und die „schlechteste" Alternative (bzw. die „beste" und „schlechteste" Ausprägung eines Attributs, s. u.) aus einer Menge von mindestens drei aussuchen [28, 29].

Entfernung zwischen Entscheidungsparametern BWS geht davon aus, dass die Probanden jede mögliche Alternative in einer Auswahlmenge prüfen und anschließend eine doppelte Wahlhandlung vornehmen, indem sie sowohl das beste, als auch das schlechteste Angebot identifizieren. Somit wählen die Probanden die Eigenschaften, Ausprägungen oder Alternativen mit der höchsten Entfernung voneinander aus. Sie legen so die maximale Distanz zwischen den Angeboten fest [30].

Falls mit den zur Wahl stehenden Entscheidungskriterien ganze Alternativen gemeint sind, handelt es sich in Abb. 1 um die Punkte B[1,2] und W[1,2], denn B[1,2] enthält von beiden gewünschten Eigenschaften mehr als alle anderen, während W[1,2] diesbezüglich von allen anderen Punkten dominiert wird.

Falls mit den zur Wahl stehenden Entscheidungskriterien dagegen Ausprägungen von Attributen gemeint sind, muss man zuerst den Wert festlegen, den das jeweils andere Attribut annehmen soll. Einfachheitshalber (und realistischerweise) soll dies hier der Wert im Status quo S sein. Entsprechend liegen die Punkte B[1] und W[1] für das Attribut a_1 auf einer Waagrechten durch S und die Punkte B[2] und W[2] für das Attribut a_2 auf einer Senkrechten durch S.

Random Utility Model Die Grundlage des Best-Worst Scalings ist die Nutzenmaximierung, welche auf der Zufallsnutzentheorie (Random Utility Theory) beruht. Diese Theorie der menschlichen Entscheidungsfindung wird auf die Arbeiten von Thurstone [31] zurückgeführt. Thurstone postulierte, dass in einem Experiment niemals alle Determinanten einer Entscheidung abgebildet werden können. Demzufolge enthalten die Entscheidungen für den Experimentator (nicht aber für die Probanden) ein Zufallselement [31]. Diese Erkenntnis führt dazu, dass sich der erreichte Nutzen aus einer deterministischen (beobachtbaren) Komponente und einer stochastischen (nicht beobachtbaren) Komponente zusammensetzt [32]. In der heutigen Terminologie

beschreibt Thurstones „law of comparative judgement“ ein Modell, das verwendet wird, um Präferenzurteile durch den Vergleich von Objekten zu erhalten. Erweitert, formalisiert und axiomisiert wurde dieses Konzept von Marschak [33] und Luce [34]. Neben dem Thurstone zugeschriebenen Probit Model wurde später von McFadden [35, 36] das sog. Multinominal Logit Model (MNL-Modell) aus der Zufallsnutzentheorie hergeleitet, das der Berechnung der Auswahlwahrscheinlichkeiten von Alternativen dient.

Fixed Utility Model Unter der Annahme, dass einerseits der Experimentator alle für die Probanden relevanten Bestimmungsgrößen der Attribute erfasst hat und andererseits die Probanden bei ihren Wahlhandlungen keine Fehler machen, sind die maximalen Distanzen zwischen „Best“ und „Worst“ feste Größen, enthalten also kein Zufallselement. Unter diesen Voraussetzungen kann man das sog. Fixed Utility Model zur Auswertung der BWS-Daten heranziehen [37]. Offenbar gilt:

$$Total(Best) \cdot Total(Worst) = r \tag{10}$$

wobei r die Zahl der vorgelegten Angebote und z. B. $Total(Best)$ die Anzahl der „Best“-Nennungen bedeutet. Aus (10) folgt nach Division durch $Total(Worst)^2$ und unter Verwendung von $Total(Worst) = r/Total(Best)$

$$\frac{Total(Best)}{Total(Wors)} = \frac{r}{Total(Worst)^2} = \frac{r}{[\frac{r}{Total(Best)}]^2} = \frac{Total(Best)^2}{r}. \tag{11}$$

Zieht man die Quadratwurzel, so erhält man

$$\sqrt{\frac{Total(Best)}{Total(Worst)}} = \frac{Total(Best)}{\sqrt{r}}. \tag{12}$$

Unter idealen Bedingungen müsste demnach die Quadratwurzel des Verhältnisses zwischen den „Best“- und „Worst“-Nennungen mit der Zahl der vorgelegten Angebote r abnehmen, und zwar nicht linear, sondern degressiv. Für eine weiterführende Darstellung sei insbesondere auf Louviere [38] und Crouch and Louviere [37] verwiesen.

In der Realität wird der Experimentator nie alle Bestimmungsgrößen des Nutzens kennen; nur schon deshalb werden ihm die Entscheidungen der Probanden ein Stück weit zufällig erscheinen. Darüber hinaus ist damit zu rechnen, dass die Probanden (wie im täglichen Leben auch) mitunter Fehler machen. Das Modell des Zufallsnutzens (Random Utility Model) berücksichtigt diese Tatsachen, indem es den Individuen nur noch die Maximierung des Nutzens im Erwartungswert (d. h. im Durchschnitt über viele Wiederholungen hinweg) unterstellt [35, 36]. Dies ist eine viel schwächere Anforderung als die traditionelle Nutzenmaximierung, weil die Probanden bei jeder einzelnen Wahlhandlung „daneben liegen“ können. Der Nutzenwert, den sie (nach ihren Optimierungsanstrengungen) tatsächlich erreichen, wird indirekter Nutzen genannt. Er ist zwar nicht beobachtbar, hängt jedoch von beobachtbaren Bestimmungsgrößen ab, nämlich den Attributen der gewählten Alternative sowie persönlichen Merkmalen wie Alter, Geschlecht und Gesundheitszustand. Im Maxdiff-Modell, das nach Flynn (2010) eine Version von BWS darstellt, besteht dann die Distanz zwischen „Best“ und „Worst“ aus einer deterministischen (erklärbaren) und einer probabilistischen (nicht erklärbaren) Komponente [27, 39–41]:

$$D_{ij} = (\gamma \cdot x_{ij}) + \varepsilon_{ij}. \tag{13}$$

Die latente (nicht beobachtbare wahre) Differenz D_{ij} zwischen den Angeboten i und j ist gleich dem Produkt des Regressionskoeffizienten des Attributs γ und dem Unterschied im Wert der Bestimmungsgrößen (x_{ij}: z. B. Distanz im Attribut) und dem nicht beobachtbaren (unerklärbaren) Anteil ε_{ij}. Wie immer in der statistischen Inferenz gilt zwar der sog. Störterm ε_{ij} als nicht beobachtbar, doch soll er einem bestimmten Verteilungsgesetz (Normalverteilung) folgen, typischerweise mit Erwartungswert Null und konstanter Varianz. Dagegen lässt sich die übliche Annahme der Unabhängigkeit über die Zeit nicht gut aufrechterhalten, da die gleiche Person im Verlauf des Experiments eine Entscheidung nach der anderen trifft. Wenn sie z. B. eine Tendenz hat, „schlechte“ Alternativen in einen Topf zu werfen, dürfte sie wiederholt Fehler bei der Identifikation der maximalen Nutzendifferenz machen [42].

Um die Parameter γ zu schätzen, welche den Einfluss der BW-Unterschiede in den Attributen abbilden, eignen sich das (multinomiale) Probit- und Logit-Verfahren (MNL). Das klassische MNL-Verfahren, welches vielfach für DCE verwendet wird, lässt sich allerdings nicht unmittelbar auf BWS übertragen, da es lediglich die Wahrscheinlichkeit der Wahl einer, d. h. der „besten“ Alternative abbildet [43]. BWS verlangt dagegen die Wahl sowohl der besten als auch der schlechtesten Alternative aus einer Auswahlmenge. Man teilt entsprechend die beiden Wahlhandlungen in zwei voneinander unabhängige Entscheidungen auf.

Unabhängigkeit der Wahlentscheidungen Die soeben beschriebene Aufspaltung hat zur Folge, dass die Annahme der vollständigen Unabhängigkeit für BWS erst recht nicht aufrechtzuerhalten ist. In Anlehnung an Marley und Louviere (2005) lassen sich vier Modifikationen des MNL-Verfahrens unterscheiden [44]: Das konsistente Best-Worst-Extremwert-Zufallsnutzenmodell, das Maxdiff-Modell, das Biased Maxdiff-Modell sowie das Konkordante Best-Worst-Modell. Die zwei erstgenannten Verfahren folgen aus der Zufallsnutzentheorie. Die beiden letztgenannten Verfahren

Abb. 2 Beispiel eines Wahlszenarios für BWS Fall 1

Wichtigstes Merkmal	Merkmale der Therapie	Unwichtigstes Merkmal
X	Einstellung des Langzeitzuckerwertes	
	Mögliche Gewichtsveränderungen	
	Risiko von Magen-Darm-Problemen	X

beruhen dagegen auf zusätzlichen Annahmen über das individuelle Wahlverhalten [40]. Zur ausführlichen Darstellung aller vier Modelle sei an dieser Stelle auf Marley und Louviere (2005) verwiesen [44].

Ableitung der Wahlwahrscheinlichkeit Ausgehend von dieser Gleichung kann das folgende Modell abgeleitet werden, welches die Wahrscheinlichkeit der Wahl des Paars ij durch ein Entscheider n darstellt:

$$P(ij/C_n) = P\left[(\gamma \cdot x_{ij} + \varepsilon_{ij}) > (\gamma \cdot x_{ik} + \varepsilon_{ik})\right], \quad \forall k,\ l \in C_n,\ ij \neq k,l. \tag{14}$$

C_n: Untergruppe der Stimuli, mit denen der Proband konfrontiert wird. $P(ij/C_n)$: Wahrscheinlichkeit, ein Paar i und j aus einer Auswahlmenge (Choice Set) zu wählen.

Wie einige Autoren aufgezeigt haben, induziert die Annahme der unabhängigen identischen Gumbel-Verteilung ein multinominales Logit-Modell (MNL) [42, 45]. Die Wahlwahrscheinlichkeiten können dann wie folgt dargestellt werden:

$$P(ij/C_n) = \frac{\exp(\delta \cdot x_{ij})}{\sum_{ik} \exp(\delta \cdot x_{ik})}, \quad \forall \delta_{ik} \in C_n. \tag{15}$$

Diese Gleichung wird folgendermaßen umgeschrieben, da die beiden Skalenwerte die Lokalität der Items auf der unterliegenden Skala repräsentieren. Daraus folgt:

$$P(ij/C_n) = \frac{\exp(\delta \cdot x_{ij} - \delta \cdot x_{ik})}{\sum_{ik} \exp(\delta \cdot x_{ij} - \delta \cdot x_{ik})}, \quad \forall \{\delta \cdot x_{ij} - \delta \cdot x_{ik}\} \in C_n. \tag{16}$$

2.3 Varianten von Best-Worst Scaling

Beim BWS können drei Varianten unterschieden werden. Alle drei Varianten zielen darauf ab, aus der beobachteten Wahlhandlung der Probanden zusätzliche Informationen zu gewinnen, um die Präferenzen mit erhöhter Genauigkeit zu erfassen. Die nachstehenden Ausführungen zeigen, inwieweit sich die drei Varianten voneinander unterscheiden und welche Stärken bzw. Schwächen sie in der Anwendung aufweisen.

2.3.1 BWS Variante 1: Object Case

Die erste Variante von BWS wird in der Literatur auch als BWS Variante 1 (oder Object Scaling) bezeichnet und stellt die Urform des BWS dar, wie sie von Finn und Louviere im Jahre 1992 vorgestellt wurde [45]. Es handelt sich hierbei um die Bestimmung der Wichtigkeit der Eigenschaften [29]. Entsprechend haben die zur Auswahl stehenden Attribute nur eine Ausprägung; sie werden auf die verschiedenen Auswahlmengen (Choice Sets) verteilt, so dass sich nur deren Zusammensetzung ändert. Die Abb. 2 illustriert das Vorgehen für drei relevante Attribute. Die befragten Personen wählen dann jeweils das am meisten und das am wenigsten präferierte Attribut des Szenarios aus [28]. Die Zahl der Wahlhandlungen ist abhängig von der Anzahl der Attribute. BWS Variante 1 stellt eine Alternative zu den traditionellen Methoden der Präferenzmessung (wie z. B. die Abfrage mittels Rating, insbesondere Likert-Skalen) dar [29].

Doch BWS Variante 1 hat den Nachteil der mangelnden Genauigkeit und Trennschärfe. Dies soll am eingangs angeführten Beispiel eines Insulinpatienten und eines OAD-Patienten gezeigt werden. Die beiden Attribute seien nach wie vor ein Niveau des Langzeit-Zuckerwertes (a_1) und ein bestimmter Grad der Gewichtsreduktion (a_2). Im Falle des Object Scaling können sie jedoch nur die Werte 0 (im Szenario nicht vorhanden) und 1 (im Szenario vorhanden) annehmen.

Die Abb. 3 enthält somit vier mögliche Alternativen, dargestellt durch die Punkte: $(0, 0)$, $(0, 1)$, $(1, 0)$, und $(1, 1)$. Es geht (genau wie in der Abb. 1) auch hier darum, die Steigung der Indifferenzkurve als Ausdruck der relativen Wichtigkeit der beiden Eigenschaften zu ermitteln. Der Punkt $(0, 0)$ als „Worst" ist nicht informativ, da durch ihn gar keine Indifferenzkurve gehen kann (jede andere Kombination der Eigenschaften ist besser). Dies gilt auch für den Punkt $(1, 1)$, denn jede andere Kombination ist schlechter.

Im Teil A der Abb. 3 handelt es sich um einen OAD-Patienten. Seine Indifferenzkurven verlaufen flach, denn das Gewicht ist relevant, so dass eine geringe Reduktion des Gewichts durch eine wesentliche Senkung des Langzeit-Zuckerwertes kompensiert werden müsste. Ihr Ausgangspunkt kann nur der Punkt $(0, 1)$ sein, denn $(1, 0)$, also ein völliger Verzicht auf Gewichtsreduktion ist für diesen Probandentyp ausgeschlossen. Doch alle Indifferenzkurven zwischen I^{I} und I^{II} kommen in Frage. Die „Best"-

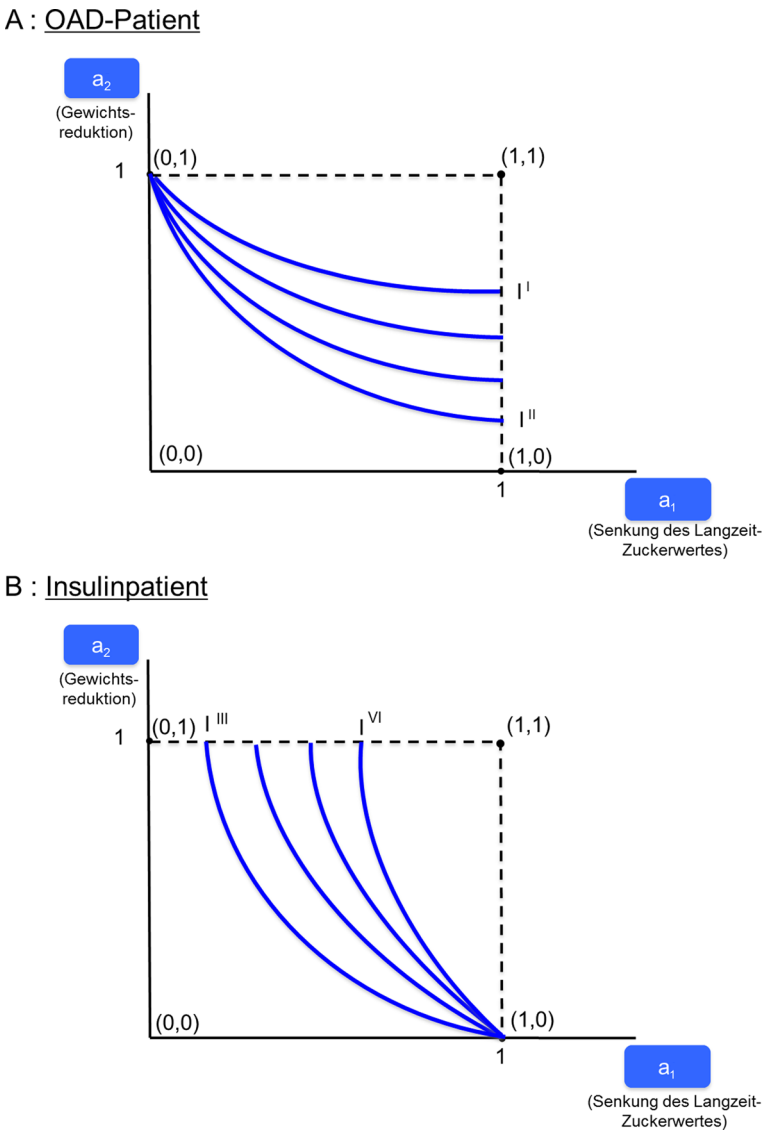

Abb. 3 Präferenzermittlung mit BWS (Object Scaling)

Nennungen können nichts anderes als $(1, 1)$ lauten, und die „Worst"-Nennungen $(0, 0)$. Damit gelingt es jedoch mit BWS nicht, die Menge der möglichen Indifferenzkurven und die Werte ihrer Steigungen einzugrenzen.

Im Teil B der Abb. 3 wird die gleiche Analyse für einen Insulinpatienten durchgeführt. Seine Indifferenzkurven verlaufen steil, weil eine geringe Senkung des Langzeit-Zuckerwertes durch eine erhebliche Reduktion des Gewichtes abgegolten werden müsste. Wieder fallen die Punkte $(0, 0)$ und $(1, 1)$ als Ausgangspunkte der Indifferenzkurven außer Betracht. Punkt $(0, 1)$ kommt dafür ebenso wenig in Frage, weil Insulinpatienten an einer Senkung des Langzeit-Zuckerwertes (a_1) sehr interessiert sind und genau dieses Attribut fehlt hier. Es verbleibt Punkt $(1, 0)$, von welchem beliebige Indifferenzkurven ausgehen, solange ihre Steigung (im Absolutwert) kleiner ist als jene des OAD-Patienten im Teil A. Die Menge der zulässigen Indifferenzkurven ist also nur durch I^{III} und I^{IV} beschränkt, und mit BWS lässt sie sich auch diesmal nicht weiter einschränken. Man sieht sofort zwei Dinge: Die jeweiligen Steigungen der Indifferenzkurven lassen sich nur höchst ungenau ermitteln, und eine Unterscheidung zwischen Insulinpatienten und OAD-Patienten ist beinahe unmöglich, weil sich die beiden zulässigen Mengen weitestgehend überdecken. Mit Object Scaling lassen sich somit die Wichtigkeit von Attributen sowie Unterschiede in dieser Wichtigkeit nur sehr ungenau ermitteln.

Dagegen werden Probleme, die bei skalenbasierten Methoden auftreten, vermieden. Skalierungsprobleme entfallen dank der auf 0–1 normierten Spannweite, so dass es zu keinen Verzerrungen von Mittelwerten und Varianzen kommt und valide Vergleiche möglich werden [46, 47]. Es kommt auch weniger zu Antworten, von denen die Probanden glauben, sie seien sozial erwünscht, weil sie Kombinationen von Attributen gegeneinander abwägen (sog. Trade-offs vornehmen und Zielkonflikte lösen) müssen [48]. Aus dem glei-

Abb. 4 Beispiel eines Wahlszenarios für BWS Variante 2

Bestes Merkmal	Merkmale der Therapie	Schlechtestes Merkmal
X	Leichte Unterzuckerung (ohne Symptome)	
	Gut eingestellter Langzeitzuckerwert (7,0-7,5%)	
	Hohes Risiko einer Genitalinfektion (20%)	
	Gewichtszunahme um 4 kg	X
	Niedriges Risiko einer Harnwegsinfektion (5%)	
	Mittleres Risiko von Magen-Darm-Problemen (8%)	

chen Grund werden Attribute auch kaum als gleich gut eingestuft (was bei Rating-Skalen der Fall sein kann). Aufgrund dieser Vorteile wird Variante 1 des BWS beispielsweise für die Einschätzung von Gesundheitszuständen bzw. der Lebensqualität oder für die Messung der Mitarbeiterzufriedenheit verwendet [49, 50].

2.3.2 *BWS Variante 2: Profile Case*

Die zweite Form des BWS ist das sogenannte Attribute oder Profile Scaling (BWS Variante 2) [51]. Im Gegensatz zu Variante 1 werden die Attribute hier über eine Spanne von Ausprägungen dargestellt, d. h. durch mehrere Ausprägungen beschrieben. Die Auswahlmengen (Choice Sets) enthalten stets die gleichen Eigenschaften, unterscheiden sich aber in deren Ausprägungen. Die befragten Personen geben an, welche Komponente der jeweiligen Auswahlmenge (Choice Sets) die beste und welche die schlechteste ist [27]. Anhand dieser Bewertungen kann aus der Summe der Teilnutzenwerte der Gesamtnutzen für jede einzelne Ausprägung bestimmt werden [44].

Die Variante 2 weist sowohl gegenüber der Variante 1 als auch gegenüber einem DCE (oder auch der Conjoint Analyse) gewisse Vorteile auf. Im Gegensatz zu Variante 1 können verschiedene Ausprägungen der Attribute bei der Bewertung berücksichtigt werden, was jede beobachtete Wahlhandlung informativer macht. Im Vergleich zur Conjoint Analyse sind Aufwand und kognitive Belastung der befragten Personen geringer, da die Probanden jeweils mit nur einer Auswahlmenge (einem konkreten Beispielfall) konfrontiert werden [46, 52]. Diese beiden Vorteile erlauben eine Erhöhung der Zahl der zu bewertenden Attribute (vgl. Abb. 4). Im Zusammenhang mit einer Diabetestherapie könnten dies die Merkmale „Mögliche Unterzuckerung“, „Einstellung des Langzeit-Zuckerwertes“, „Risiko einer Genitalinfektion“, „Mögliche Gewichtsveränderung“, „Risiko einer Harnwegsinfektion“ und „Risiko von Magen-Darm-Problemen“ sein.

Aus Sicht der (mikroökonomischen) Nutzentheorie ist allerdings die Variante 2 (Profile Case) des BWS mit starken Annahmen belastet [30, 52]. Auch mit diesem Verfahren zielt man darauf ab, über die Steigung der Indifferenzkurve die relative Wichtigkeit der Eigenschaften auszudrücken. In einem bestimmten Szenario erweise sich für einen OAD-Patienten B[1] als die beste Ausprägung von a_1, W[1] dagegen als die schlechteste (vgl. Teil A der Abb. 5). Es handelt sich bei diesem Beispiel um eine ceteris-paribus-Wertung, d.h. alle anderen Eigenschaften werden auf einem bestimmten Wert konstant gehalten. Grundsätzlich müsste dieser Wert im Experiment explizit vorgegeben werden (der Einfachheit halber soll der Wert hier auf dem Status-quo-Wert gehalten sein). Entsprechend liegen W[1] und W[2] auf einer Waagrechten durch den Punkt S. Die wahre relative Wichtigkeit der beiden Attribute wird durch den Winkel α angezeigt, der angibt, welche Verschlechterung des Langzeit-Zuckerwertes der Proband akzeptieren würde, um in den Genuss einer Gewichtsreduktion zu kommen.

Dieser Winkel wird jedoch in Variante 2 des BWS anders bestimmt, nämlich als Verhältnis der Maxdiff-Werte, gegeben durch $\{B[2]-W[2]\}/\{B[1]-W[1]\}$ (im Teil A der Abb. 5 ist die vertikale Differenz $\{B[2]-W[2]\}$ zuerst von S weg verschoben worden). Dieses Verhältnis ergibt den (Tangens des) Winkel(s) $\tilde{\alpha}$, der in Abb. 5 erheblich vom wahren Winkel α abweicht. Man würde fälschlicherweise auf eine recht hohe Bedeutung des Langzeit-Zuckerwertes bei OAD-Patienten schließen, weil eine diesbezügliche Einbuße offenbar durch eine wesentlich erhöhte Gewichtsreduktion kompensiert werden müsste.

Darüber hinaus wird auch der Vergleich zwischen OAD-Patienten und Insulinpatienten in BWS Variante 2 verfälscht. Im Teil B der Abb. 5 verläuft die Indifferenzkurve steil, weil Insulinpatienten besonderen Wert auf den Langzeit-Zuckerwert legen (vgl. auch Abb. 1). Der abgebildete Insulinpatient identifiziert die gleichen „Best“- und „Worst“-Werte wie der OAD-Patient für Attribut a_1 (gegeben a_2 ist auf dem Status-quo-Wert) und a_2 (gegeben a_1 ist auf seinem Status-quo-Wert). Die BWS-Wertungen ergeben den Winkel $\tilde{\beta}$, der (zufällig) der wahren Steigung β der Indifferenzkurve in der Umgebung des Punktes S entspricht. Nun aber sind die Nutzenwerte und ihre Differenzen zwischen Individuen nicht unbedingt vergleichbar. Beispielsweise könnte die Distanz zwischen B[2] und W[2] für einen

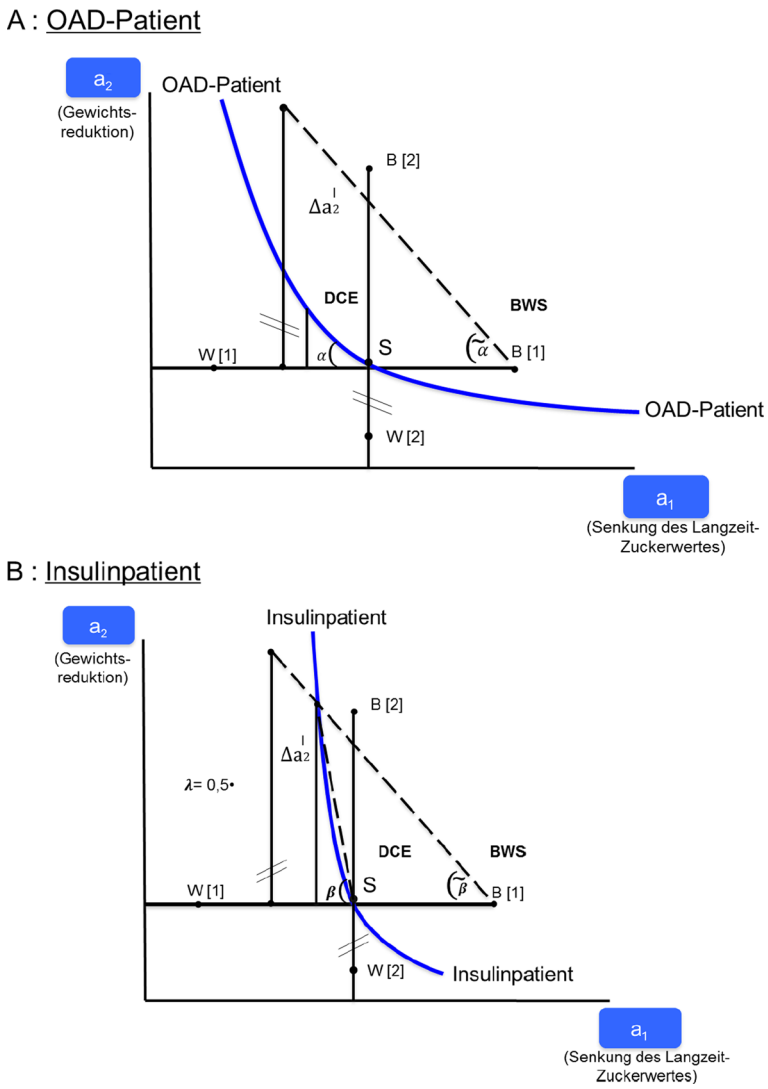

Abb. 5 Präferenzermittlung mit BWS (Profile Case)

Insulinpatienten mit einem anderen Nutzenunterschied verbunden sein als für einen OAD-Patienten; dies würde einer Transformation $\varphi(\cdot)$ des Nutzens von U zu $\tilde{U}$ entsprechen, wie in Abschn. 2.1 diskutiert wurde. Die Nutzendifferenz sei hier für den OAD-Patienten nur halb so groß ($\lambda = 0{,}5$ bei Homothetik). Aus dem gemessenen Winkel $\tilde{\beta}$ wird dann $0{,}5\beta$ – weit geringer als β. Aus dem Vergleich von $0{,}5\tilde{\beta}$ und $\tilde{\alpha} < 0{,}5\tilde{\beta}$ des Teils A der Abb. 5 würde man den Schluss ziehen, der Langzeit-Zuckerwert sei den Insulinpatienten weniger wichtig als den OAD-Patienten. Der Vergleich der Indifferenzkurven zeigt jedoch, dass dies nicht zutrifft. Offensichtlich ist bei BWS Variante 2 eine (nicht erkannte) Heterogenität der Präferenzen ein großes, nicht zu lösendes Problem.

2.3.3 BWS Variante 3: Multiprofile Case

Die dritte Variante des BWS ist der sogenannte Multiprofile Case [29, 53]. Es handelt sich hierbei um eine Sonderform des Discrete-Choice Experiments (DCE), welche in der (mikroökonomischen) Nutzentheorie verankert ist.

Die BWS Variante 3 beruht (analog zum DCE) auf den ökonometrischen Modellen von McFadden [36] und der Nachfragetheorie von Lancaster [8]. Aufgrund der mikroökonomischen Fundierung, insbesondere der Zufallsnutzentheorie (Random-Utility Theory, RUT), ist eine wohlfahrtstheoretische Interpretation der Ergebnisse möglich [42, 54].

Im Unterschied zu den zuvor beschriebenen Formen wählen die Probanden bei BWS Variante 3 zwischen ganzen Alternativen (beschrieben durch unterschiedlichen Ausprägungen der Eigenschaften) innerhalb einer Auswahlmenge (Choice Set). Die Szenarien unterscheiden sich demnach nicht durch die Attribute, sondern durch deren jeweilige Ausprägungen. Dies entspricht einem Best-Worst Discrete-Choice Experiment (BWDCE). BWDCE generiert mehr Informationen aus einem einzigen Wahlszenario als das klassische DCE, indem einmal mehr nicht nur nach der besten

Abb. 6 Beispiel eines Wahlszenarios für BWS Variante 3

Merkmal	Alternative 1 (Therapie 1)	Alternative 2 (Therapie 2)	Alternative 3 (Therapie 3)
Mögliche Unterzuckerung	Schwer (starke Symptome)	Mäßig (mäßige Symptome)	Leicht (ohne Symptome)
Einstellung des Langzeitzuckerwertes	Sehr gut (6,5 – 7,0 %)	Gut (7,0 – 7,5 %)	Weniger gut (7,5 – 8,0 %)
Risiko einer Genitalinfektion	Mittel 15 von 100 (15 %)	Niedrig 10 von 100 (10 %)	Hoch 20 von 100 (25 %)
Mögliche Gewichtsveränderungen	+ 4 kg	– 4 kg	+/– 0 kg
Risiko einer Harnwegsinfektion	Niedrig 5 von 100 (5 %)	Hoch 10 von 100 (10 %)	Mittel 7,5 von 100 (7,5 %)
Risiko von Magen-Darm-Problemen	Hoch 12 von 100 (12 %)	Mittel 8 von 100 (8 %)	Niedrig 4 von 100 (4 %)
Beste Alternative		X	
Schlechteste Alternative	X		

(bzw. der am meisten präferierten), sondern zugleich auch nach der schlechtesten (bzw. der am wenigsten präferierten) Alternative gefragt wird (vgl. Abb. 6).

Bislang wurde die dritte Variante des BWS nur sehr vereinzelt zur Präferenzmessung im Gesundheitswesen herangezogen. Die geringe Unterstützung durch Standardsoftware, die zur Erstellung entsprechender Fragebögen herangezogen werden kann, mag ein Grund dafür sein [46]. Einige Erhebungen konnten jedoch zeigen, dass die gewonnenen Ergebnisse ebenso reliabel sind wie vergleichbare Ergebnisse von konventionellen DCE-Analysen [55]. Dies war zu erwarten, weil ja beide Verfahren in der (mikroökonomischen) Nutzentheorie verankert sind und Skalen-invariante Aussagen ermöglichen. Darüber hinaus lassen sich mit Experimenten vom Typ BWS Variante 3 Präferenzen genauer messen als mit DCE oder direktem Abfragen der Zahlungsbereitschaft (Willingness-to-pay) [52]. Auch dies entspricht den Erwartungen, soll doch mit der Frage nach der besten und der schlechtesten Alternative eine größere Informationsmenge (bei gleicher Anzahl der Choice Sets) gewonnen werden [52]. Eine andere Untersuchung zeigt dagegen, dass das DCE gegenüber der BWS Variante 3 einige Vorteile aufweist. So zeigen die Ergebnisse der Studie, dass die Bewertungsaufgaben beim DCE für die Probanden einfacher und schneller zu bewältigen sind. Zudem sind die Ergebnisse in Hinblick auf die Konsistenz und Genauigkeit zuverlässiger [7].

3 Design und Analyse

3.1 Erhebungsdesigns

Beim Erhebungsdesign handelt es sich um die Konstruktion zu bewertender Entscheidungsszenarien aus den Kombinationen von Eigenschaften (für Variante 1) und/oder ihren Ausprägungen (für Variante 2 und 3). Wie im Falle der DCE gibt es für BWS-Studien eine Reihe von Möglichkeiten. Ein sog. „Full-Factorial"-Design kommt höchstens für maximal drei Attribute mit je drei Ausprägungen in Frage, da dann die Zahl der Szenarien bereits $3^3 = 27$ beträgt. In allen übrigen Fällen muss auf ein sog. fraktionell-faktorielles Design zurückgegriffen werden. Die Auswahl der Szenarien, die jeweils das Maximum an Information generiert, hängt u.a. von den Beziehungen der Attribute untereinander ab [56]. Drei Vorgehensweisen werden hier näher beschrieben, wobei keine den anderen in allen Aspekten überlegen ist [57].

3.1.1 Manuelle Designs

Aus einer vollständigen Auflistung der Möglichkeiten lassen sich geeignete Designs manuell erstellen, indem man auf eine ausgeglichene Zahl hoher und niedriger (vermuteter) Nutzenwerte, möglichst geringe Korrelation der Attribute (Orthogonalität), ausgewogene Repräsentation und minimale Überlappung der Ausprägungen achtet [58]. Ist die so reduzierte Zahl der von den Probanden verlangten Wahlhandlungen immer noch zu groß, so bildet man sog. Designblöcke. Beispielsweise wird die Hälfte der Probanden nur mit Block Nr. 1 konfrontiert, die andere Hälfte mit Block Nr. 2, wobei es von Vorteil ist, die Zuordnung zufällig zu gestalten, um Verzerrungen durch nichtrepräsentative Stichproben zu vermeiden. Ein häufig im Rahmen von Best-Worst Scaling angewendetes Verfahren ist das sog. Balanced Incomplete Block Design (BIBD). In einem BIBD (v, b, r, μ, λ) haben die Entscheidungsszenarien jeweils $v \geq 2$ Attribute und sind in $b > 0$ Blöcken zusammengefasst, wobei jeder Block genau μ Choice Sets umfasst (Symmetrieanforderung; zudem gilt $v > \mu > 0$). Zudem ist jedes Szenario in $r > 0$ Blöcken enthalten und jedes Wahlszenario tritt simultan in genau $\lambda > 0$ Blöcken auf. Da λ konstant gehalten wird, ist das

Design balanciert: da $\mu < v$ (d. h. kein Block enthält alle Entscheidungsszenarien) ist es unvollständig [59].

Für eine effiziente Designkonstruktion eines BWS-Experimentes gelten die folgenden zwei Gleichungen, mittels derer die einzelnen Elemente zu kombinieren sind:

$$b \cdot \mu = v \cdot r, \tag{17}$$

$$\lambda \cdot (v - 1) = r \cdot (\mu - 1). \tag{18}$$

Da ein BIBD die Symmetrieanforderung erfüllen muss, ist die Anzahl konstruierbarer BIBDs begrenzt. Allerdings finden sich in der Literatur verschiedene Übersichten zur Erstellung, Berechnung und Operationalisierung dieser manuellen Designform. An dieser Stelle sei vor allem auf die Autoren Cochran und Cox (1992) verwiesen, die eine Vielzahl möglicher BIBDs erstellt haben, die beliebig Anwendung finden können [60]. Daneben existieren eine Reihe weiterer Ansätze zur Erstellung effizienter Designs; es sei auf Chrzan und Orme [57] sowie auf Louviere und Hensher [42] verwiesen. Darüber hinaus wurden in den letzten Jahren auch einige optimale und quasi-optimale (near-optimal) Designs entwickelt, welche der manuellen Designkonstruktion dienen [61].

3.1.2 Optimierte Designs

Statt ein Design mühsam von Hand zu entwickeln, bedient man sich häufig computergestützter, automatisierter Verfahren. So kann beispielsweise mittels des Softwarepakets SAS aus einer Vielzahl möglicher Designs mit Hilfe verschiedener Suchalgorithmen das effizienteste Design ermittelt werden [62]. Aber auch einfache orthogonale Main-Effect-Designpläne (OMEP) stehen (z. B. in SPSS) für die Erstellung eines BWS-Designs zur Verfügung. Sie sind einfach in der Handhabung und kommen beim BWS bevorzugt zum Einsatz. Es sei hier auf die Website von Sloane verwiesen, wo entsprechende Designs zu finden sind [63]. Allerdings haben OMEP den Nachteil, dass die sog. Grenzrate der Substitution (die Steigung der Indifferenzkurve in Abb. 1) nicht vom Niveau der Attribute abhängig gemacht werden kann. Dies widerspricht der Konvexität der Indifferenzkurve, die anzeigt, dass eine Verschlechterung des Langzeit-Zuckerwertes nur wenig durch ein anderes Attribut kompensiert zu werden braucht, wenn sie im Status quo weitgehend gegeben ist, doch massiv kompensiert werden muss, sobald sie im Status quo bereits stark eingeschränkt ist (rechts oder links von S). Diese Annahme erscheint sinnvoll, denn ein Weniger von einem Attribut bedeutet wohl einen geringen Verlust, so lange man damit gut versorgt ist. In der zu schätzenden Nutzenfunktion $U = f(a_{1,2})$ verlangt dies bei einer linearen Spezifikation Interaktionsterme von der Art $(a_1 \cdot a_2)$,

$$U = \gamma_0 + \gamma_1 \cdot a_1 + \gamma_2 \cdot a_2 + \gamma_3 (a_1 \cdot a_2) + \varepsilon, \tag{19}$$

so dass (ausgewertet am Erwartungswert des Störterms, d. h. $E(\varepsilon) = 0$)

$$\partial U / \partial a_1 = \gamma_1 + \gamma_3 a_2 \tag{20}$$

sowie

$$\partial U / \partial a_2 = \gamma_2 + \gamma_3 a_1. \tag{21}$$

Dies bedeutet, dass der Grenznutzen eines Attributs jeweils vom Wert des anderen Attributs (allgemeiner: der anderen Attribute) abhängt. Dieser Zusammenhang lässt sich bei einem orthogonalen Design nicht abbilden; allerdings sind die Stichproben vielfach zu klein, als dass man statistisch zwischen $a_{1,2}$ und $(a_1 \cdot a_2)$ unterscheiden könnte, so dass γ_1 und γ_2, nicht aber γ_3 signifikant von Null verschieden geschätzt werden (Problem der Multikollinearität). Dennoch ist es angezeigt, sich bei der Hypothesenbildung zu fragen, ob es Gründe für solche Interaktionen der Attribute gibt und gegebenenfalls vom orthogonalen Design abzuweichen.

3.2 Datenanalyse

Auch für die Analyse der mit BWS gewonnenen Daten stehen verschiedene Verfahren zur Auswahl.

3.2.1 Count-Analyse

Orthogonale BWS-Designs lassen sich mittels der sogenannten Count-Analyse auswerten, bei der lediglich die Wahlhäufigkeiten ausgezählt werden. Somit kann dieses Verfahren sowohl auf aggregierter Ebene (über alle befragten Personen) als auch auf individueller Ebene (für einen Probanden) angewandt werden [18, 29]. Der Best-Worst-Score ist durch die Differenz $(Total(Best) - Total(Worst))$ definiert [18, 64]. Er erlaubt Aussagen über die Wichtigkeit und Rangordnung der Attribute, aber nicht im Sinne der Grenzrate der Substitution (vgl. Abb. 1), da die Distanz zwischen „Best“ und „Worst“ nicht Skalen-invariant ist. Andere Autoren schlagen eine Verhältnis- statt der Differenzenbildung vor, indem sie Total(Best) durch Total(Worst) dividieren und die Quadratwurzel ziehen, analog zur (7), sei es auf der Ebene des einzelnen Attributs oder auf der Ebene ganzer Entscheidungsszenarien [27]. Durch die Standardisierung der Best-Worst-Scores soll eine studienübergreifende Vergleichbarkeit der Ergebnisse gewährleistet werden. Hierzu wird der errechnete Best-Worst-Score durch das Produkt aus Häufigkeit des Auftretens der einzelnen Eigenschaftskriterien (Eigenschaft, Ausprägung, Alternative) und der Größe der Stichprobe (Anzahl der Probanden) dividiert [18].

Allerdings ist festzuhalten, dass mit diesen Mittelwertbildungen kein Rückschluss auf die relative Wichtigkeit von Attributen im Sinne der ökonomischen Analyse (Grenzrate

der Substitution) ermöglicht wird. Denn die subjektive Distanz zwischen „best“ und „worst“ dürfte je nach Studie unterschiedlich ausfallen (analog zur Abb. 3). Dies hat jedoch zur Folge, dass beispielweise die Frage „Gibt es Unterschiede in der Abwägung zwischen Nebenwirkung und Verlängerung der Lebensdauer zwischen jungen und alten Menschen?“ nicht beantwortet werden kann.

3.2.2 Multinomial Logit

Im BWS wird erhoben, ob ein Attribut, eine Ausprägung oder eine Alternative als wichtigstes (analog zu „best“) oder unwichtigstes („worst“) Merkmal genannt wurde. Dies verlangt eine doppelte Kodierung, nämlich best $= 1$, falls das Attribut als das wichtigste in der Kombination erscheint, und best $= 0$ sonst, sowie worst $= 1$, falls es als das unwichtigste erscheint, und worst $= 0$ sonst. Die beiden zu analysierenden Variablen nehmen im Ergebnis lediglich die Werte 0 und 1 an. Eine Wahrscheinlichkeit kann nur zwischen 0 und 1 variieren, so dass mit dem Logit-Verfahren Tendenzen (Wahrscheinlichkeiten) geschätzt werden, dass ein Attribut in der betrachteten Kombination vorkommt.

Eine lineare Regression würde ebenfalls zu Werten zur Darstellung der relativen Wichtigkeit führen, die jedoch außerhalb dieses zulässigen Bereichs von Null und Eins liegen (also nicht als Wahlwahrscheinlichkeiten interpretiert werden können). Manche andere setzen sich über diese Einschränkung hinweg und verwenden die Methode der gewichteten Kleinsten Quadrate (die Gewichtung wird nötig, weil die $(0, 1)$-Eigenschaft der zu erklärenden Variablen zu einer Störgröße ε führt, die nicht, wie im Grundmodell verlangt, eine konstante Varianz hat) [43].

Die Logit-Transformation transformiert den $(0, 1)$-Bereich in einen $(-\infty, +\infty)$-Bereich, der mit der Methode der Kleinsten Quadrate analysiert werden kann. Der Preis für dieses Vorgehen besteht darin, dass sich die Logit-Koeffizienten nicht unmittelbar als Unterschiede in der Wahrscheinlichkeit interpretieren lassen, sondern ihrerseits einer Transformation unterworfen werden müssen. Da eine Regression stets den bedingten Erwartungswert der zu erklärenden Größe ermittelt, kann sie die Präferenzmessung für eine einzelne Person nicht leisten. Die Schätzung der Nutzenfunktion erfolgt somit lediglich auf aggregierter und nicht auf individueller Ebene [65]. Dies ist aber in vielen Anwendungen von BWS auch nicht notwendig; überdies können mit Hilfe von Interaktionstermen (s. o.) stets sozioökonomische Eigenschaften der Probanden mit berücksichtigt werden, was gruppenspezifische Aussagen erlaubt. Detaillierte Darstellungen finden sich bei Flynn et al. (2008) und Wirth (2010) [40, 65].

3.2.3 Latent Class Analyse

Die Latent Class Analyse ist auch als Clustering bekannt; sie bietet sich insbesondere dann an, wenn es nicht gelingt, mit Hilfe beobachtbarer sozioökonomischer Eigenschaften homogene Gruppen zu bilden [66]. Beispielsweise können die geschätzten Werte der marginalen Zahlungsbereitschaft innerhalb einer bestimmten Altersklasse besonders stark streuen, was auf eine versteckte Heterogenität schließen lässt. Es besteht dann die Vermutung, dass es latente Unterschiede im Wahlverhalten gibt, die nichts mit dem Alter zu tun haben. Das MNL-Verfahren muss dann so verallgemeinert werden, dass es aus den Beobachtungen Rückschlüsse auf zwei oder mehr latente Gruppen erlaubt, deren Umfang nicht bekannt ist. Entsprechend werden zusammen mit der Wahrscheinlichkeit, dass ein Proband einem bestimmten Segment angehört, segmentspezifische Nutzenfunktionen ermittelt, ohne die Stichprobe aufzuspalten. Der individuelle Nutzenwert einer Alternative lässt sich dann errechnen aus dem segmentspezifischen Schätzwert, gewichtet mit der Wahrscheinlichkeit der Zugehörigkeit zu dieser Gruppe [67]. Da diese Wahrscheinlichkeit von der vermuteten Zahl der latenten Gruppen abhängt und zudem im Zuge der Schätzung immer wieder neu bestimmt wird, müssen sehr viele Beobachtungen vorliegen, um zu statistisch signifikanten Ergebnissen zu gelangen. Bei dem nachstehend beschriebenen hierarchischen Bayes-Modell kommt man mit kleineren Stichproben aus, muss dafür aber einschränkende Annahmen einführen [46, 64].

3.2.4 Hierarchisches Bayes-Modell

Die Hierarchischen Bayes-Modelle (HB-Modelle) werden zunehmend für die Analyse von DCEs verwendet [68]. Sie folgenden dem Bayesianischen Ansatz, indem sie von einer Priori-Verteilung der zu schätzenden Parameter ausgehen und auf Grund der beobachteten Daten eine Posteriori-Verteilung schätzen. So lässt sich vorhandenes Wissen berücksichtigen, z. B. ob ein bestimmtes Attribut positiv oder negativ bewertet wird (was z. B. bei einem Preisattribut immer angenommen werden kann). Als Priori-Verteilung wird in der Regel die multinominale Normalverteilung unterstellt, deren Symmetrieeigenschaft jedoch nicht immer geeignet ist. Dank der Unterstellung einer spezifischen Priori-Verteilung lassen sich aus HB-Modellen auch bei geringer Antwortzahl pro Proband stabile individuelle Best-Worst-Werte herleiten. Das Verfahren ist auch insofern effizient, als für die Nutzenschätzung eines einzelnen Probanden die Wahlhandlungen aller befragten Personen mit einbezogen werden [69]. Eine MNL-Schätzung bestimmt zudem die individuelle Wahrscheinlichkeit, mit der ein bestimmtes Entscheidungsszenario gewählt wird. Das Vorgehen bei der Auswertung von BWS-Daten ähnelt stark dem Vorgehen bei der HB-Schätzung eines DCE (Choice-based Conjoint Analysis), mit dem einen Unterschied, dass im Falle von BWS die Wahl der schlechtesten Alternative zusätzlich zu analysieren ist. Für die Bestimmung der Posteriori-Verteilung

steht keine geschlossene Lösung zur Verfügung, sondern sie erfolgt mit Hilfe der sog. Markov-Chain-Monte-Carlo-Simulation, die z. B. durch Sawtooth Software unterstützt wird [24, 40].

4 Abschließende Beurteilung

BWS hat einen breiten Anwendungsbereich, der von der Schätzung von Nutzen und Werten der marginalen Zahlungsbereitschaft für spezifische Eigenschaften und ganze Alternativen bis zur Voraussage und Akzeptanz innovativer Gesundheitsgüter und -leistungen reicht. In den Managementwissenschaften und dem Marketing hat sich BWS bereits gut etabliert, in der Gesundheitsökonomie und der Versorgungsforschung hingegen deutlich weniger, allerdings mit zunehmender Tendenz [23].

Flynn et al. (2008) weiteten ihre Untersuchung der Patientenpräferenzen in Bezug auf dermatologische Beratung zu einem Methodenvergleich aus [65]. Sie stellten das Verfahren der gewichteten Kleinsten Quadrate dem MNL-Verfahren gegenüber und konnten ein hohes Maß an Übereinstimmung feststellen. Sie behaupten überdies, eine differenzierte Zuordnung von Ausprägungen zu einer latenten Nutzenskala sei mit traditionellen DCEs nicht möglich [65]. Dies sei ein großer Mangel, denn genau die Unterschiede im Niveau eines Attributs spielten doch in der Gesundheitspolitik eine große Rolle (beispielsweise der Unterschied zwischen gar keinen und positiven Wartezeiten). Überdies müsste ein solche „marginale“ Veränderung mit einer fundamentalen Veränderung, bei der ein Attribut dazukommt (oder wegfällt) verglichen werden können. Einmal mehr stellt sich allerdings die Frage, weshalb mit Hilfe von BWS eine Abbildung auf einer Nutzenskala im Gegensatz zu DCE möglich sein soll, wo doch BWS (insbesondere Variante 3) lediglich eine Verfeinerung des DCE-Verfahrens darstellt, das genauere (aber nicht grundsätzlich andere) Messungen von Nutzendifferenzen erlaubt. Diese Einschätzung wird auch durch die Studie von Chrzan und Golovashkina (2006) bestätigt, bei der sechs Verfahren zur Ermittlung der Wichtigkeit von Attributen verglichen werden. Die Autoren kommen zum Schluss, dass BWS sowohl am besten zwischen den Attributen differenziert als auch tatsächliche Entscheidungen am besten voraussagt [70].

Die wesentliche Stärke von BWS besteht in der besseren Informationsausbeute, indem die Probanden veranlasst werden, nicht nur eine Aussage (Alternative 1 ist besser als Alternative 2), sondern zwei Aussagen zu machen (Alternative 1 ist „Best“, Alternative 2 ist „Worst“). Die Präferenzstruktur der befragten Personen lässt sich so präziser bzw. gleich präzise, doch mit geringerem Stichprobenumfang ermitteln. Zudem erlauben wiederholte BWS-Fragen, durch den schrittweisen Ausschluss von vorher als „Best“ oder „Worst“ identifizierter Alternativen vollständige statt nur partielle Rangfolgen herzuleiten. Als weiterer Vorteil von BWS wird die verringerte kognitive Belastung der Probanden angeführt; ob dies tatsächlich zutrifft, ist allerdings nicht abschließend geklärt und sollte in weiteren Studien überprüft werden (siehe hierzu: Severin, Schmidtke [71]). Immerhin scheint es den Probanden leichter zu fallen, die beiden jeweiligen Extrempunkte auf ihrer Nutzenskala zu bestimmen, als in komplexen Entscheidungsszenarien zwischen zwei oder mehreren Alternativen mit vielen Attributen die am meisten präferierte Alternative auszuwählen oder gar eine vollständige Rangfolge bei einer Vielzahl von Attributen zu bestimmen [45]. Diese Aussage bezieht sich allerdings auf BWS Variante 2 (Profile Case), dem die wichtige Eigenschaft der Skalen-Invarianz fehlt (vgl. Abschn. 2.3.2). Der gleiche Vorbehalt ist an der Behauptung anzubringen, BWS ermögliche die Identifikation individueller Präferenzskalen. Was als Vorteil bleibt, ist einmal mehr die genauere Messung durch verbesserte Ausbeute der Information.

Als Verfeinerung von DCE hat das BWS schließlich Vorteile im Vergleich zu herkömmlichen Messskalen zur Erhebung von Präferenzen (wie das Rating, das Ranking oder auch die Punkteverteilung) [23]. Doch diese Vorteile gehen auf die nutzentheoretische Verankerung der DCE zurück, die dafür sorgt, dass die Probanden bei ihrer Entscheidungsfindung zwischen den Attributen abwägen, genau wie im täglichen Leben, wo man die Vor- und Nachteile einer Alternative gewichten und gegeneinander aufrechnen muss (vgl. Abschn. 2.1).

Neben Vorteilen weist BWS auch Schwächen auf, die letztlich auf die einfache Tatsache zurückgehen, dass auch in Experimenten zusätzliche Informationen nur zu erhöhten Kosten zu haben sind. So steigt für die Probanden der für die Auswahlentscheidung erforderliche zeitliche Aufwand [66, 70]. Offenbar trifft das Argument der kognitiven Einfachheit von BWS [27] in der Realität nicht unbedingt zu. Darüber hinaus können sich die Befragten nicht wie beim Rating an einer vorgegebenen Skala orientieren, sondern unterliegen einem Entscheidungszwang [72]. Dieser Einwand trifft allerdings sämtliche Verfahren, welche den Probanden eine Entscheidung abverlangen (wobei sich diese Nachteile der Wahlhandlung auch im täglichen Leben nicht vermeiden lassen). Die mittels BWS zusätzlich gewonnene Information ist (zumindest in Variante 2) nicht so wertvoll wie von einigen Autoren angeführt. Denn die Abfrage der besten und schlechtesten Ausprägungen ergibt noch keine Informationen über die Attraktivität der jeweiligen Auswahlmenge selbst. Somit können keine Rückschlüsse auf das effektive Inanspruchnahme- bzw. Nachfrageverhalten der Patienten und Konsumenten gezogen werden. Erachtet der Befragte beispielsweise alle Optionen der Wahlentscheidung als unwichtig oder schlecht bzw. wichtig oder

gut, hat er keine Möglichkeit, dies in der Befragung zum Ausdruck zu bringen. Eine Lösung besteht im Hinzufügen einer entsprechenden Ablehnungsoption (z. B. die Frage: „Würden Sie dieses Szenario im Vergleich zu Ihrem Status Quo akzeptieren bzw. annehmen?“) [43]. Schließlich ist auch die statistische Auswertung der durch BWS gewonnenen Daten aufwändiger als im Falle eines DCE. Schon beim DCE gilt es zu berücksichtigen, dass die gleiche Person immer wieder die Wahlhandlung vornimmt, mit der Folge, dass die Störterme ε nicht unabhängig über die Zeit „aus der Urne gezogen“ sein können. Diese sog. Autokorrelation der Residuen führt dazu, dass die Standardfehler der geschätzten Effekte grösser sind als bei Unabhängigkeit. Übertragen auf das BWS stellt sich die Frage, ob sich die Autokorrelation nur je auf ε_B (die „best-Residuen“ untereinander) und auf ε_W (die „worst-Residuen“ untereinander) beziehen soll. Möglicherweise hat der Umstand, dass ein Proband in der letzten Wahlentscheidung einem Attribut, einer Ausprägung oder einer Alternative einen geringeren Nutzenwert als erwartet zuordnet ($\varepsilon_W < 0$) Konsequenzen für die Zuordnung in der jetzigen Entscheidung, z. B. in der Form $\varepsilon_B > 0$, was negative Autokorrelation „übers Kreuz“ bedeuten würde.

5 Ausblick

Studien zeigen, dass das BWS unabhängig vom Design und Stichprobenumfang ebenso reliable Resultate liefern kann wie traditionelle Formen der Discrete-Choice-Experimente [28, 53]. Am überzeugendsten erscheint BWS als Verfeinerung der herkömmlichen DCE (hier die BWS Variante 3). In dieser Funktion eröffnet BWS neue Optionen im Bereich der Gesundheits- und Versorgungsforschung. Eine solche Option ist die genauere Erfassung der Heterogenität von Präferenzen [6].

Es gibt jedoch auch einige Unklarheiten, die im Rahmen künftiger Arbeiten analysiert werden sollten. Flynn und Louviere [43] zufolge fehlen Richtlinien zur Festlegung der Stichprobengröße für BWS-Studien. Offen ist auch die adäquate Modellierung der Zufallskomponente der Nutzenfunktion (Random Utility Component), wenn sich die Heterogenität der Präferenzen im „Best“ und „Worst“ niederschlagen. Es fragt sich auch, ob zur Berücksichtigung von sozioökonomischen Eigenschaften gleich vorgegangen werden kann wie in DCE, wo Interaktionsterme (wie in Gleichung (19)) eingefügt werden. Es wäre denkbar, dass „Best“-Antworten anders von Alter, Geschlecht und insbesondere Einkommen abhängen als „Worst“-Einschätzungen. Allgemein wäre für die Gesundheitspolitik von Interesse zu erfahren, ob die gemessenen Prioritäten der Befragten mit ihren soziodemographischen Merkmalen in Verbindung stehen [29]. Als zusätzliche Komplikation könnte die Wertung eines Attributs vom Niveau anderer Attribute abhängen, wie dies grundsätzlich mit Blick auf die Steigung der Indifferenzkurven in Abb. 1 zu erwarten wäre. Solche Abhängigkeiten wurden bislang wenig geprüft, vor allem weil die Stichproben zu klein waren, um die entsprechenden Koeffizienten (γ_3 in Gleichung (19)) genügend genau schätzen zu können. Dank der verbesserten Informationsausbeute mittels BWS könnte sich dies in Zukunft ändern.

Abschließend bleibt festzuhalten, dass Forschungsbedarf vor allem auch hinsichtlich der Implementierung von BWS in die gängige Praxis der Experimente besteht. Bislang sind die hier diskutierten Methoden Medizinern, Versorgungsforschern und anderen (gesundheitspolitischen) Entscheidungsträgern unzureichend bekannt, so dass ein großer Nachholbedarf besteht. Für die Entscheidungsfindung im Gesundheitswesen sind neben Expertenurteilen vermehrt (subjektive) Wertungen der Betroffenen notwendig, die nur mittels Präferenzmessungen zur Verfügung gestellt werden können.

Literatur

1. Holcombe R. The median voter model in public choice theory. Public Choice. 1989;61(2):115–25.
2. Buchanan JM, Tollison RD. The theory of public choice, II. Ann Arbor: University of Michigan Press; 1984.
3. Liebl A. Insulintherapie bei Typ-2-Diabetes. Diabetologe. 2007; 3:221–32.
4. Mühlbacher A, et al. Patients preferences regarding the treatment of type II diabetes mellitus: comparison of best-worst scaling and analytic hierarchy process. Value in Health. 2013;16(7); A446.
5. Berner S, Leukert K, Zweifel P. Präferenzen für Krankenversicherung in Deutschland und den Niederlanden (Preferences for Health Insurance in Germany and the Netherlands: A Two-country Study). In: Franz W, et al, Hrsg. Experimentelle Wirtschaftsforschung, Wirtschaftswissenschaftliches Seminar Ottobeuren. Tübingen: Siebeck; 2009. S. 125–45.
6. Gelhorn H. Preferences for medication attributes among patients with type 2 diabetes mellitus in the UK. Diabetes Obes Metab. 2013;15:802–9.
7. Xie F, et al. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best–worst scaling? Eur J Health Econ. 2012;15(3):1–8.
8. Lancaster K. Consumer demand: a new approach. New York: Columbia University Press; 1971.
9. Nida-Rümelin J. Entscheidungstheorie und Ethik. München: Utz; 2005. S. 406.
10. Backhaus K, Lütgemüller F, Weddeling M. Messung von Kundenpräferenzen für produktbegleitende Dienstleistungen. ServPay Arbeitspapier, Working paper; 2007 (1).
11. Kockelman KM, Krishnamurthy S. A new approach for travel demand modeling: linking Roy's identity to discrete choice. Transp Res, Part B, Methodol. 2004;38(5):459–75.

12. Sattler H. Methoden zur Messung von Präferenzen für Innovationen. Schmalenbach Z Betriebswirtsch Forsch. 2006;54(06):2006.
13. Merino-Castello A. Eliciting consumers preferences using stated preference discrete choice models: contingent ranking versus choice experiment. UPF economics and business working paper; 2003 (705).
14. Bateman IJ, et al. Economic valuation with stated preference techniques: a manual. Cheltenham Glos.: Edward Elgar; 2002.
15. Helm R, Steiner M. Präferenzmessung: Methodengestützte Entwicklung zielgruppenspezifischer Produktinnovationen. Stuttgart: W. Kohlhammer Verlag; 2008.
16. Klein M. Die Conjoint-Analyse: Eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen. 2002.
17. Schöffski O, von der Schulenburg J-MG. Gesundheitsökonomische Evaluationen. 4. Aufl. Heidelberg: Springer; 2012.
18. Cohen E. Applying best-worst scaling to wine marketing. Int J Wine Bus Res. 2009;21(1):8–23.
19. Baumgartner H, Steenkamp JBEM. Response styles in marketing research: a cross-national investigation. J Mark Res. 2001;38:143–56.
20. Sato Y. How to measure human perception in survey questionnaires. Int J Anal Hier Process. 2009;1(2):64–82.
21. Alwin DF, Krosnick JA. The measurement of values in surveys: a comparison of ratings and rankings. Public Opin Q. 1985;49(4):535–52.
22. Stallmeier C. Die Bedeutung der Datenerhebungsmethode und des Untersuchungsdesigns für die Ergebnisstabilität der Conjoint-Analyse. Dissertation, Roderer Verlag; 1993 (75), S. 87–90.
23. Simon A. Patienteninvolvement und Informationspräferenzen zur Krankenhausqualität. Der Unfallchirurg. 2011;114(1):73–8.
24. Weinert R. Eigentum als eine Determinante des Konsumentenverhaltens: Das Beispiel Zweitwohnung (Universität St. Gallen). Göttingen: Cuvillier; 2010.
25. Paulhus DL. Measurement and control of response bias. In: Robinson JP, Shaver PR, Wrigthsman LS, Hrsg. Measures of personality and social psychological attitudes. San Diego: Academic Press; 1991. S. 17–59.
26. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica: Journal of the Econometric Society. 1979;47(2):263–91.
27. Flynn TN. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Rev Pharmacoecon Outcomes Res. 2010;10(3):259–67.
28. Lancsar E, Louviere J. Estimating individual level discrete choice models and welfare measures using best-worst choice experiments and sequential best-worst MNL. University of Technology, Centre for the Study of Choice (Censoc); 2008, S. 1–24.
29. Louviere JJ, Flynn TN. Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia. Patient. 2010;3(4):275–83.
30. Marley AAJ. The best-worst method for the study of preferences: theory and application. Working paper, Department of Psychology, University of Victoria Victoria, Canada; 2009.
31. Thurstone LL. A law of comparative judgment. Psychol Rev. 1927;34(4):273.
32. Hensher DA, Rose JM, Greene WH. Applied choice analysis: a primer. Cambridge: Cambridge University Press; 2005.
33. Marschak J. Binary-choice constraints and random utility indicators. In: Proceedings of a symposium on mathematical methods in the social sciences, Cowles foundation discussion papers; 1960.
34. Luce RD. Individual choice behavior a theoretical analysis. New York: Wiley; 1959.
35. McFadden D. The choice theory approach to market research. Mark Sci. 1986;5(4):275–97.
36. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, Hrsg. Frontiers in econometrics. New York: Academic Press; 1974.
37. Crouch GI, Louviere JJ. International convention site selection: a further analysis of factor importance using best-worst scaling. Queensland: CRC for Sustainable Tourism; 2007.
38. Louviere JJ. Best–worst scaling. In: Workshop on theory and example applications. Sydney: School of Marketing at the University of Technology in Sydney, Australia; 2006.
39. Hall J, et al. What influences participation in genetic carrier testing? Results from a discrete choice experiment. J Health Econ. 2006;25(3):520–37.
40. Wirth R. Best–Worst Choice-Based Conjoint-Analyse: Eine neue Variante der wahlbasierten Conjoint-Analyse. Marburg: Tectum-Verlag; 2010.
41. Mühlbacher A, Bethge S, Tockhorn A. Präferenzmessung im Gesundheitswesen: Grundlagen von Discrete-Choice-Experimenten. Gesundh.ökon Qual.manag. 2013;18(4):159–72.
42. Louviere JJ, Hensher DA, Swait JD. Stated choice methods: analysis and applications. Cambridge: Cambridge University Press; 2000.
43. Flynn TN, et al. Best–worst scaling: what it can do for health care research and how to do it. J Health Econ. 2007;26(1):171–89.
44. Marley AAJ, Louviere JJ. Some probabilistic models of best, worst, and best–worst choices. J Math Psychol. 2005;49(6):464–80.
45. Finn A, Louviere JJ. Determining the appropriate response to evidence of public concern: the case of food safety. Journal of Public Policy & Marketing. 1992. 12–25.
46. Kübler RV; Best/worst scaling. In: Albers SK, Konradt UW, Wolf J, Hrsg. Methodik der empirischen Forschung. Wiesbaden: Gabler; 2013.
47. Cohen S, Orme B. What's your preference? Mark Res. 2004;16:32–7.
48. Auger P, Devinney TM, Louviere JJ. Using best-worst scaling methodology to investigate consumer ethical beliefs across countries. J Bus Ethics. 2007;70(3):299–326.
49. Lee JA, Soutar GN, Louviere J. Measuring values using best-worst scaling: the LOV example. Psychol Mark. 2007;24(12):1043–58.
50. Garver MS, Williams Z, LeMay SA. Measuring the importance of attributes in logistics research. Int J Logist Manage. 2010;21(1):22–44.
51. Marley AAJ, Flynn TN, Louviere JJ. Probabilistic models of set-dependent and attribute-level best-worst choice. J Math Psychol. 2008;52(5):281–96.
52. Louviere JJ, Islam T. A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best–worst scaling. J Bus Res. 2008;61(9):903–11.
53. Marti J. A best-worst scaling survey of adolescents' level of concern for health and non-health consequences of smoking. Soc Sci Med. 2012;75(1):87–97.
54. Gerard K, Shanahan M, Louviere J. Using stated preference discrete choice modelling to inform health care decision-making: a pilot study of breast screening participation. Appl Econ. 2003;35(9):1073–85.
55. Louviere JJ, et al. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. J Choice Model. 2008;1(1):128–63.
56. Johnson RM, Orme BK. How many questions should you ask in choice-based conjoint studies. In: Conference proceedings of the ART forum, Beaver Creek. 1996.
57. Chrzan K, Orme B. An overview and comparison of design strategies for choice-based conjoint analysis. Sawtooth Software Research Paper Series. 2000.
58. Huber J, Zwerina K. The importance of utility balance in efficient choice designs. J Mark Res. 1996;33:307–17.

59. Smith NF, Street DJ. The use of balanced incomplete block designs in designing randomized response surveys. Aust N Z J Stat. 2003;45(2):181–94.
60. Cochran WG, Cox GM. Experimental designs. 2. Aufl. New York: Wiley; 1992.
61. Burgess L, Street DJ. Optimal designs for choice experiments with asymmetric attributes. J Stat Plan Inference. 2005;134(1):288–301.
62. Kuhfeld WF. Marketing research methods in SAS: experimental design, choice, conjoint, and graphical techniques. Cary, NC, SAS-Institute TS-722; 2009.
63. Sloane NJ. A library of orthogonal arrays. 2006 [cited 2005]; Available from: http://neilsloane.com/oadir/index.html.
64. Coltman TR, Devinney TM, Keating BW. Best–worst scaling approach to predict customer choice for 3PL services. J Bus Logist. 2011;32(2):139–52.
65. Flynn TN, et al. Estimating preferences for a dermatology consultation using best–worst scaling: comparison of various methods of analysis. BMC Med Res Methodol. 2008;8(1):76.
66. Cohen S. Maximum difference scaling: improved measures of importance and preference for segmentation. In: Sawtooth software conference proceedings, Sequim, WA. 2003.
67. Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutchen AL, Hrsg. Applied latent class analysis. Cambridge: Cambridge University Press; 2002. S. 89–106.
68. Train KE. Discrete choice methods with simulation. Cambridge: Cambridge University Press; 2002.
69. Hartmann A, Sattler H. Wie robust sind Methoden zur Präferenzmessung? Universität Hamburg, Fachbereich Wirtschaftswissenschaft, Institut für Handel und Marketing; 2002.
70. Chrzan K, Golovashkina N. An empirical test of six stated importance measures. Int J Mark Res. 2006;48(6):717–40.
71. Severin F, et al. Eliciting preferences for priority setting in genetic testing: a pilot study comparing best-worst scaling and discrete-choice experiments. Eur J Hum Genet. 2013;21(11):1202–8.
72. Bacon L, et al. Comparing apples to oranges. Mark Res. 2008;38(2):143–56.